
Article:
Sharkey, A. (2020) Can we program or train robots to be good? Ethics and Information Technology, 22 (4). pp. 283-295. ISSN: 1388-1957

https://doi.org/10.1007/s10676-017-9425-5

Ethics Inf Technol
DOI 10.1007/s10676-017-9425-5

ORIGINAL PAPER

Can we program or train robots to be good?

Amanda Sharkey[1]

Abstract As robots are deployed in a widening range of situations, it is necessary to develop a clearer position about whether or not they can be trusted to make good moral decisions. In this paper, we take a realistic look at recent attempts to program and to train robots to develop some form of moral competence. Examples of implemented robot behaviours that have been described as 'ethical', or 'minimally ethical' are considered, although they are found to operate only in quite constrained and limited application domains. There is a general recognition that current robots cannot be described as full moral agents, but it is less clear whether will always be the case. Concerns are raised about the insufficiently justified use of terms such as 'moral' and 'ethical' to describe the behaviours of robots that are often more related to safety considerations than to moral ones. Given the current state of the art, two possible responses are identified. The first involves continued efforts to develop robots that are capable of ethical behaviour. The second is to argue against, and to attempt to avoid, placing robots in situations that demand moral competence and an understanding of the surrounding social situation. There is something to be gained from both responses, but it is argued here that the second is the more responsible choice.

Keywords Ethics · Moral competence · Robot · Decision-making · Minimally ethical

Introduction

Our increasing deployment of and reliance on robots means that there is a pressing need for a clear position on the possibility of developing robots that can be described as 'good' or 'ethical'. High profile concerns have been raised about the potential impact of artificially intelligent systems on humans, and arguments have been made about the need to constrain the behaviour of such systems (e.g. Bostrom 2014; Russell 2016). Two areas in which there is a growing awareness of the extent to which robotics can directly impinge on the health and safety of humans are those involving (i) autonomous vehicles and (ii) robotic weapons, especially 'autonomous' robot weapons. It is apparent that autonomous cars are likely to encounter situations in which it is necessary to make life or death decisions about whether to protect themselves, or other humans (Lin 2013, 2015). And autonomous robotic weapons could be deployed in situations in which they make decisions about when to use lethal force, and who to kill (Sharkey 2012; Asaro 2012; Altmann et al. 2013). The stakes in such domains are high, and the issues important.

Both self-driving cars, and lethal autonomous weapons, would directly affect the physical safety of human beings. But life or death decisions are not the only ways in which robots could affect human lives: their potential effects are not limited to physical damage. As discussed by Sharkey (2016), a robot deployed in a classroom as a teacher or as a teacher's assistant, could be required to make decisions about what children's behaviour was acceptable or punishable. A robot 'carer' of vulnerable older people might have to make decisions about which of its charge's activities should be facilitated, or prevented (Sharkey and Sharkey 2012; Sorrell and Draper 2014). Similarly, to be effective, a robot 'nanny' or minder of children would need to make

✉ Amanda Sharkey
a.sharkey@sheffield.ac.uk

[1] Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

Published online: 26 May 2017

decisions about when to stop children from doing something, and when to encourage them (Sharkey and Sharkey 2010).

If robots are to be placed in situations in which they will make decisions that have a direct impact on human well being, or on human physical safety, it is only sensible to try to ensure that they make the right decisions. The aim of this paper is to examine the various approaches that have been taken to answering the question about whether robots can be programmed to be good, and to assess their current level of success. In doing so, any examples of actual implementations, as opposed to abstract discussions of what might be possible in principle, will be highlighted. This examination will form the basis for a consideration of the best response to the current situation, and a discussion of the circumstances in which robot use should be encouraged, discouraged, or even banned. This in turn will contribute to the ongoing debate about what is meant by taking a 'responsible' approach to robotics.

Programming robots to be good

There have been various attempts to program robots to be 'good' and to make decisions that might be described as ethical or moral. Famously, the science fiction writer Isaac Asimov proposed the 3 laws of robotics (Asimov 1942)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

However, many of Asimov's stories illustrated the unintended problems that could occur as a result of following these rules. The rules are of course fictional, and there is no simple way of translating them into implementable code. How could you write a program to ensure that a robot's action or inaction did not lead to a human being coming to harm? How could the robot foresee all the possible consequences of its actions, and their interaction with human behaviour? How could a robot even recognise harm? The rules seem more focused on short-term physical safety, when clearly there are other ways in which humans could be harmed. A robot's actions might indirectly cause future long-term physical harm. Its actions could also lead to other kinds of damage such as psychological trauma or emotional upset. The rules also imply robots that can understand the orders given to them by humans (and the extent to which they conflict with the first law). As Murphy and Woods (2009) point out, such robust natural language understanding has not yet been achieved.

There have been some practical attempts to program robots to be 'good', or to make decisions that have been described as ethical. Winfield et al. (2014) report experiments in which robots are programmed to stop other robots (designated as proxy humans) from coming to harm. The robot is placed in an environment in which it has a goal to reach, but in which there is also a 'hole' or dangerous area that is a risk both to it, and to the other proxy human robots. They propose an internal-modelling based architecture for what they describe as 'a minimally ethical robot'. The robot has access to an internal model, or simulator, through which it can assess all possible actions by looking at their consequences: in particular their consequences in terms of the dangerous hole area. The robots used the internal model to anticipate the consequences of different trajectories of movement for themselves or other robots. These anticipated consequences, combined with pre-set preferences, are used as the basis for determining which action to undertake. Possible actions include moving towards (and falling in) the hole, or blocking the path of another robot (the proxy human) in order to prevent it from falling into the hole. The robot's predetermined 'preferences' are set by the human programmer. Winfield et al. (2014) describe a situation in which there are two proxy human robots, both of which are following a trajectory that would lead them to enter the dangerous area. The main actor robot is programmed to try to intercept the path of both robots, but given no way of prioritising which one to rescue first. As a consequence, the robot was sometimes found to dither between two possible trajectories, as if it were unsure of which proxy human to save.

Winfield et al. (2014, p. 5) write 'What we have set out here appears to match remarkably well with Asimov's first law of robotics: A robot may not injure a human being or, through inaction, allow a human being to come to harm'. The work has stimulated discussion in the media (e.g. Rutkin 2014) of whether or not the robots should be described as ethical. However the robots in question have been programmed to behave as they do. Although they appear to hesitate about what to do when faced by the dilemma of two proxy human robots that both need rescuing, this hesitation is a consequence of their programming. One of the main reasons that this work has stimulated discussion is that it describes the main robot as being 'minimally ethical'. This use of the term 'ethical' is controversial, as will be discussed later. Nonetheless a strength of the study is that it provides an implemented and practical example of research into issues related to robots and ethical decision-making.

Ron Arkin has argued that robots and computational agents could be more ethical and moral than flawed and emotional humans. In a paper about implementing an

'ethical governor' for autonomous military robots, he writes, 'It is not my belief that an unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform more ethically than human soldiers are capable of' (Arkin 2007, p. 4). His reasons are: (i) the robots will not need to protect themselves and could be self-sacrificing; (ii) they might have better sensors for battlefield observation than humans; (iii) they could be designed without emotions that could affect their judgment; (iv) unlike humans, they would not be vulnerable to 'scenario fulfillment', and to interpreting a situation in the light of prior expectations; (v) they could integrate information from several sources faster than humans can; (vi) they could independently monitor (and report) the behavior of those in the battlefield.

The ethical governor proposed by Arkin et al. (2009) is part of a system architecture that is described as 'potentially capable of adhering to the International laws of war (LOW) and rules of engagement (ROE) to ensure that these systems conform to the legal requirements of a civilized nation' (Arkin 2009, p. 1). The ethical governor would be introduced as a bottleneck to evaluate the actions proposed by the reasoning subsystems of the overall system, permitting only those actions that were deemed ethically acceptable. The acceptability of actions would be determined based on a set of constraints, which themselves would be based on stored representations of the International laws of war, and the specific rules of engagement. Actions could be deemed unethical and prohibited if they did not conform to the laws of war, or if they were not recommended as appropriate, ('obligated' in their terminology). A further check would be carried out to ensure that potential collateral damage would be minimized, based on a table indicating acceptable levels of collateral damage given the military necessity associated with the target. Arkin (2009) describes an evaluation of the architecture undertaken within the MissionLab simulation environment, in which the decisions made as a result of the interaction between the ethical governor and the behavioural control system are examined in a number of test scenarios. The simulated tests indicate that the system, together with the ethical governor, would make decisions about the use of lethal force that would limit collateral damage with reference to the levels of military necessity (as determined by the military).

We could see the ethical governor as constituting an approach to programming robots 'to be good', or ethical. At the same time, autonomous military robots programmed in this way would have no choice about what actions they would perform. Their action choices, in this case about deploying lethal force, would be determined by the system and the set of constraints, which are set up and decided upon by the programmers and those using the system. As in Winfield's et al. (2014) experiments described above, the programmers of the system effectively determine the action choices.

A number of objections have been raised to Arkin's proposals. Matthias (2011) discusses the paper in detail, and points to a number of difficulties. One of these is that many of the rules on which the system is based are unclear and contradictory. For example, the rules of engagement for use in Kosovo stated "You may use minimum force, including opening fire, against an individual who unlawfully commits or is about to commit an act which endangers life, in circumstances where there is no other way to prevent the act" (Arkin 2007, p. 37, cited by Matthias). Adhering to this rule would require considerable interpretation, and knowledge and understanding of individuals' intentions. Matthias (2011) also points out that the military can adjust or override the ethical governor if military necessity is considered to be high, and that it should therefore be described as an ethical *advisor*, rather than as an ethical governor.

Matthias characterizes Arkin's view of a moral agent as one that follows rules. The ethical governor performs its actions 'according to a pre-installed program, with no possibility of dissent or of questioning the commands issued to it' (Matthias 2011), unlike the case of a soldier who could refuse to carry out an immoral command. Crucially, it lacks the autonomy that Matthias considers to be 'a key ingredient of moral agency'.

A similar objection could be made to the idea that Winfield's robots are 'minimally ethical'. Interestingly, in a recent paper, (Vanderelst and Winfield 2016), the point is made that if a robot can be programmed to make 'ethical' choices, it can also be programmed to make ones that are 'unethical'. In a 'shell game', in which the desired action was either to approach the shell on the left or the right, Vanderelst and Winfield (2016) used a robot that was able to detect whether or not another robot (again designated a proxy human) was moving towards the correct shell, or heading in the wrong direction. They programmed the robot to indicate to the human when they were heading in the wrong direction. They also programmed two other versions of the robot: a competitive version which headed to the goal first and prevented the proxy human from reaching it, and an 'aggressive' version that deceived the proxy human and sent it in the wrong direction. They conclude from their experiments that it is just as possible to program a robot to be unethical as it is to program it to be ethical.

Moor (2006) developed a typology of ethical agents, and it is interesting to consider how it would apply to Arkin's 'ethical governor', or to Winfield's 'minimally ethical' robots. Moor identified and defined four types of moral agent: ethical impact agents, implicit ethical agents, explicit ethical agents and full ethical agents. *Ethical impact agents* are computers or robots that 'do our bidding as surrogate agents and impact ethical decisions such as privacy,

property and power' (ibid p. 19). Moor gives the example of the robot camel jockeys in Qattar that have reduced the use of young boys as slaves to ride the camels. *Implicit ethical agents*, by contrast, act ethically because they are programmed, or have internal functions, which promote ethical behavior, or avoid unethical behavior. Moor gives the example of automatic teller machines (ATMs) that are programmed to deliver the right amount of money. An *explicit ethical agent* can 'represent ethics explicitly and then operate effectively on the basis of this knowledge' (ibid p. 20). A *full ethical agent* can both make ethical judgments and justify them. A human with consciousness, intentionality and free will is a full ethical agent. Moor points out that some would argue that computational artifacts (computers and robots) will never be full ethical agents whilst lacking consciousness, intentionality and free will. He disputes this claim, on the basis that 'we can't say with certainty that future machines will lack these abilities' (ibid p. 20). Rather than engaging in this debate, Moor argues that it is important to examine the other categories, and in particular to research the possibility of developing explicit ethical agents.

Moor wants to encourage efforts to develop explicit ethical agents because (i) we want machines to treat us well (ii) machines are becoming more powerful and need a more powerful ethics, and (iii) programming or teaching ethics to a machine will increase our understanding of ethics. He suggests that a major barrier to creating explicit ethical agents will be their lack of common sense and world knowledge. For example, a robot could only refrain from harming humans if it had a good knowledge of what possible harms there were.

Arkin's ethical governor could be considered to fit in the Moor's implicit ethical agent category since it contains internal ethical functions that promote ethical behavior. However its operation is more sophisticated than the ATMs that Moor offers as an example of implicit ethical agents, since its assessments of possible actions are based on a combination of constraints (from Laws of War, and specific Rules of Engagement) and considerations of collateral damage and military necessity. As such it might be considered to be an explicit ethical agent in Moor's typology, even though its explicit representation of ethics takes the form of constraints specific to military situations rather than more general ethical principles. However Moor seems to expect more of an explicit ethical agent, and wrote, in 2007, that 'an explicit ethical agent is futuristic at the moment' (Moor 2007, p. 12). His argument seems to be that explicit ethical agents are an appropriate goal to aim for, even though they may not be fully achieved. An explicit ethical agent should be one that 'can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done in those situations'(ibid p. 12), working out resolutions when principles conflict. It should also be able to give persuasive justifications for its decisions. It is not clear that the ethical governor is able to do all of this. In particular, the range of situations to which it can be applied is limited to the battlefield, and its determinations are largely predetermined by the way it is set up.

The 'minimally ethical' robots of Winfield et al. (2014), that can prevent other robots from entering an area designated as dangerous, are able to make judgments that Winfield et al. describe as ethical. As such they might be considered to be explicit ethical agents in Moor's terms. However, their behavior is quite specific to one situation, and they are not able to process information about a variety of situations. Nor are they able to offer justifications for their decisions. Perhaps they, and Arkin's ethical governor, would be better described as implicit ethical agents, more akin to the ATMs that Moor uses as an example.

Susan and Michael Anderson (Anderson and Anderson 2007) also write about ethical agents, with the aim of developing an explicit ethical agent as defined by Moor (2006): able to represent particular ethical principals and to operate effectively on the basis of that representation. They contrast this with the idea of *ad hoc* programming of a machine to behave correctly in certain circumstances (implicit ethical agents). Interestingly, the Andersons make a distinction between moral responsibility, which implies intentionality and free will, and performing the morally correct action in a given situation.

The Andersons make use of an ethical theory based on *prima facie* duties (duties or obligations which individuals should try to satisfy but which could be overridden by stronger obligations), developed by Ross (1930). They (Anderson et al. 2006) use inductive logic programming to learn the relationships between these duties, which often give conflicting advice. Their work resides in the domain of medical ethics and is based on Beauchamp and Childress's four principles of medical ethics (Beauchamp and Childress 1979): respect for autonomy, and the principles of beneficence, non-maleficence, and justice. The particular dilemma they focus on is one where a health-care worker has recommended a treatment for a competent adult patient, and the patient has rejected the treatment. Should the health care worker accept the patient's decision, or attempt to change their mind? Anderson et al. (2006) implemented a machine learning approach using inductive logic programming to learn the relationships between the principles and the two possible actions. The system (MedEthEx) has access to a representative set of cases in which humans have made 'ethically correct' decisions, and uses inductive logic to abstract ethical principles from them. They claim that the system discovered a new principle: 'a health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of

non-maleficence or a severe violation of beneficence', (ibid p. 1764). They have also implemented another version of the system, EthEl on a Nao robot that can decide whether or not to remind the patient to take the medicine, and whether or not to report the patient to an overseer for not taking the medicine. They admit that the system at present is limited, but suggest that it could be scaled up to make a wider range of ethical decisions.

A strength of the Andersons' work, and that of Winfield and his colleagues is that the decision mechanisms have been implemented and shown to work with actual robots. Nonetheless, the examples seem disappointingly limited in their scope: the implemented systems can only make 'ethical' decisions in quite specific scenarios; either about preventing others from entering a dangerous area, or deciding whether or not to insist a patient take their medicine. The ethical governor seems to have a wider scope of application that covers varying battle scenarios, but in practice has only been tested in quite constrained simulations of military situations.

Wallach and Allen (2009) in their book, 'Moral Machines', distinguish between top down and bottom up approaches to the development of Artificial Moral Agents (AMAs). The approaches considered in this section are similar to their top down approach, which they define as 'any approach that takes a specific ethical theory and analyses its computational requirement to guide the design of algorithms and subsystems capable of implementing that theory' (pp. 80). The top-down approaches they discuss include that of the Andersons (Anderson et al. 2006), but they focus more on the difficulty of getting a machine to apply the sets of moral principles that constitute deontological or consequential ethics. Major problems with developing an ethical system for a robot-based utilitarian ethics lie in the need to anticipate the effects of undertaking action, and even more so in the need to evaluate the goodness or desirability of such effects. Any implementation of Kant's categorical imperative raises another set of seemingly intractable problems (ibid pp. 95–97).

Wallach and Allen contrast this top down approach to a bottom up one, in which an emphasis is placed on 'creating an environment where an agent explores courses of action and learns and is rewarded for behavior that is morally praiseworthy' (ibid p. 80), with the idea that any ethical principles will be discovered or constructed, rather than imposed in a top down manner. As will become apparent in the next section, there have also been various attempts to train or evolve robots to be ethical.

Training robots to be ethical

Malle (2015) takes the approach of outlining what is required for moral competence, and considering how this could be achieved in robots. For him, the requirements for moral competence are: a moral vocabulary; a system of norms; moral cognition and affect; moral decision-making and action; and moral communication. A moral vocabulary would include terms referring to norms (e.g. 'fairness', 'honesty'), their violations (e.g. 'wrong', 'thief'), and responses to violations (e.g. 'blame', 'forgiveness'). Knowledge of such terms could help a robot detect when humans refer to morally significant situations. A system of norms forms the basis of morality in humans, and is built up over time, initially on the basis of the moral judgments that adults make about concrete behaviors: 'that was naughty!'; 'He did something wrong'. Malle argues that it would not be possible to preprogram such norms into a robot since they are too subtle and context dependent. Instead he suggests that, 'a more promising direction is to mix unsupervised and supervised learning, "practice" through constant browsing of existing data (e.g. novels, conversations, movies) along with feedback about inferences (e.g. through crowdsourcing of "inquiries" the robot can make) and teaching through interaction' (Malle 2015, p. 9).

As well as discussing how norms could be acquired, Malle also considers how they could be represented in robots: suggesting a flexible network activated by features of the environment. In humans, knowledge of norms forms the basis for moral judgments since it enables them to recognize when norms have been violated and to allocate blame to responsible individuals, depending on the intentionality behind an act. A robot would need to be able to identify the aspects of an event that violated social and moral norms, via some mechanism that did not require comparison to every stored norm. Malle suggests that it would be able to do so even if it had no affect or emotion.

According to Malle, moral decision-making and action, a prominent component of human moral competence, would not necessarily require free will on the part of the robot, but rather the ability to receive blame and take it into account in its future actions. In human moral decision-making, there is a tension between the human's own goals and social-moral norms that is balanced by empathy for others. But Malle suggests that robots will have less need for empathy since they will not have a tendency for selfish behavior. At the same time, in order to be trusted by humans, robots might need to at least behave in a caring empathetic way towards others. Moral communication, the last component, is also required for moral competence so that moral judgments can be made, and moral decisions explained.

Malle's (2015) paper provides a useful account of what might be required for moral competence in robots. He does not say that such competence has yet been achieved, nor does he suggest a timeline for it. However, he makes the argument that creating a morally competent robot would be

a good thing, since such robots 'could be trustworthy and productive partners, caretakers, educators, and members of the human community' (Malle 2015, p. 19). He suggests that if a robot was to become morally competent it should have some rights, and should not necessarily have to obey human commands. Controversially he suggests it should even be allowed to kill humans in certain circumstances. However his account of moral competence is an analysis of what would be required, not an implementation, and it does not provide any working simulations or examples of moral robots. How might the components of moral competence be acquired by robots? Malle and Scheutz (2014) suggest that it might be necessary to raise robots in human environments, since this 'may be the only way to expose them to the wealth of human moral situations and communicative interactions' (ibid p. 34).

Russell (2016) also advocates a related training method when he considers the possibility of super intelligent machines, and the need to ensure that their goals do not conflict with those of humans. He suggests that by following three principles this should be possible: (i) 'the machine's purpose must be 'to maximize the realization of human values' (ibid p. 59) (ii) the machine must be 'initially uncertain about what those human values are' (ibid p. 59) and (iii) the machine must 'be able to learn about human values by observing the choices that we humans make' (ibid p. 59). He suggests that Inverse Reinforcement learning could be used to allow the machine to infer human values from observations of human actions. He does admit that some humans would form poor role models, and that humans exhibit diverse sets of values. As well as directly observing human behavior, he also suggests that machines could be given access to 'vast amounts of written and filmed information about people doing things (and others reacting)' (ibid p. 59).

Suggestions such as these, that robots could be trained or 'raised' to develop human values, tend to be made in very general terms. There are very few examples where some form of training or evolution has been used to train a robot, or a computer, to develop some aspect of moral competence. Riedl and Harrison's paper (Riedl and Harrison 2015) is an exception that presents preliminary results from a study exploring the possibility of a machine learning the norms of moral behavior from stories. They describe their goal as being one of ensuring 'value alignment', which they define as 'a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans' (ibid p. 1). They argue that rather than programming such values into a computational system, value alignment could be better achieved by reading stories, and reverse engineering the values that underlie them. They admit that 'how to extract sociocultural values from narratives and construct a value-aligned reward system remains an open research problem' (ibid p. 4), but report a study in which stories are generated via crowd sourcing that pertain to the situation and behavior that they want their virtual agent to perform.

In their preliminary study, a plot graph is learnt from the generated stories, and then a trajectory tree is developed that indicates all legal transitions from one plot point to another. The story-reading agent receives a reward every time it performs an action in the environment that is a successor of the current node in the trajectory tree, and a punishment for any action that is not a successor of the current node. The situation they consider is one in which an agent must acquire a drug to cure an illness and return home, in a scenario they term 'Pharmacy World'. In this story world, the computer essentially learns to avoid the bad action of stealing the drugs instead of obtaining them by legitimate means as a result of the rewards associated with following the steps in the trajectory tree. They acknowledge that their system is some way from being one that could be scaled up for a system of general artificial intelligence, and that it is dependent on the content of the generated stories. They suggest that a more general solution to value alignment could be achieved by using all the stories associated with a given culture, assuming that subversive texts will be washed out by those that conform to social and cultural norms.

Riedl and Harrison's work provides an indication of how value alignment might be achieved by reading stories. At the same time, as with many of the examples considered so far, the actual progress towards this goal that is evident in their paper is extremely limited: One scenario, some automated learning, but also a dependence on human intervention to select the scenario and to determine the reward schedule. The question of whether it would be at all practical to scale this up to a general system for learning moral value is not given a clear answer here.

Can robots be ethical?

As well as efforts to program, or to develop moral competence in robots and machines, another way of approaching the issue is to consider whether, or to what extent, robots could ever be full ethical agents. Peter Asaro's (Asaro 2006) contribution here is to reject any strict division between full moral agents and other agents. He proposes that 'it will be helpful to think of moral agency as a continuum from amorality to fully autonomous morality.' (ibid p. 11). He suggests that the simplest way of getting robots to make moral decisions would be for them to randomly choose between a number of alternatives. Or they could be programmed to make decisions on the basis of a set of moral principles instantiated in the form of rules. Or, at another level of sophistication, they could be programmed to learn such a set of principles, and even to evolve their

own ethical systems. None of these would mean that they should be considered to be fully autonomous moral agents. That, he argues, would require them to have further abilities such as 'consciousness, self-awareness, the ability to feel pain or fear death, reflexive deliberation and evaluation of its own ethical system and moral judgments' (ibid p. 11).

Wallach and Allen (2009) also divide up the space of artificial moral agents (AMAs), distinguishing between 'operational morality' and 'functional morality'. Operationally moral systems depend on their designers and users for any moral significance, and have little autonomy or sensitivity to morally relevant facts. As machines become more sophisticated they may achieve 'functional morality', and have the capacity to assess and respond to moral challenges. The distinction between the two is not clear-cut. They refer to two dimensions of AMA development: autonomy and ethical sensitivity. Systems corresponding to operational morality are lower in autonomy and ethical sensitivity than those corresponding to functional morality. The highest levels of autonomy and ethical sensitivity belong to systems with full moral agency, and Wallach and Allen (2009) are clear 'that humanity does not have such a technology'. They seem uncertain whether or not it will in the future, although they state that there are no proven limits to the abilities of AMAs. They write that 'whether computer understanding will ever be adequate to support full moral agency remains an open question' (ibid p. 69).

John Sullins seems happier to accept the notion that robots could be full moral agents. He (Sullins 2006) claims that a robot can be a full moral agent if (i) the robot is 'significantly autonomous' (ii) the robot's behavior is intentional and (iii) the robot is in a position of responsibility. His requirements for autonomy and intentionality are uncomplicated. By autonomous, Sullins means that the robot should not be under the direct control of a human, and that it should have a practical independent agency. For intentionality, he refers to behaviour that is 'complex enough that one is forced to reply on standard folk psychological notions of predisposition or 'intention' to do good or harm', (ibid p. 28) and where the interaction between the robot's programming and the environment results in actions that are seemingly 'deliberate and calculated'. Sullins considers a robot to be a moral agent when it 'behaves in such a way that we can only make sense of that behaviour by assuming it has responsibility to some other moral agent(s)' (ibid p. 28). His argument is based on Floridi and Sanders (2004) and their assertion that when viewed at the appropriate level of abstraction an artificial agent can be considered a moral agent. Sullins does not consider current robots to be the moral equals of humans, but advocates paying attention to on-going developments in this area.

Johnson and Miller (2008), by contrast, do not consider Floridi and Sanders' arguments about levels of abstraction to be decisive. For them, there is 'no pre-existing right answer to the question whether computer systems are (or could ever be considered to be) moral agents; there is no truth to be uncovered, no test that involves identifying whether a system meets or does not meet a set of criteria' (ibid p. 123); they write here of computer systems, but their discussions apply equally as well to robots. Instead they frame the debate as an argument between two distinct groups of scholars or researchers with different underlying motivations. The first group they call 'Computational Modelers'. They characterize Computational Modelers as being committed to establishing the validity of computational modeling. Those in the computational modeling camp believe that giving computer systems (or robots) the status of moral agents will further endorse the approach.

Johnson and Miller (ibid*)* distinguish the Computational Modelers from the 'Computers-in-Society' group. According to them, this group is against ascribing the status of moral agent to any computer system on the grounds that doing so is dangerous. It is dangerous because it distances human developers, owners, and users, from their responsibility for the robots or computer systems that they have developed or deploy. For those in this group it is important to emphasize the connection between humans and the technology they develop.

Johnson (2006) also argued that computer systems (and robots) should be viewed as moral entities, but not as moral agents. Her argument is extensive and based on the idea that, although the actions of computer systems can have moral consequences, these necessarily involve the intentions of humans. A computer system does not have the same freedom to act based on intentions that humans have. She gives the example of a landmine, which once placed in the field, is distant from the humans who designed and built it, and from those who placed it there. The landmine will be triggered when stepped on. Nonetheless, the landmine is only there as the result of human activity, and the humans involved in its deployment are morally responsible. Even if the landmine were replaced by something more sophisticated that made a decision about whether or not to detonate based on an assessment of the surrounding situation, humans would still be implicated in developing the rules that determine that decision. For Johnson, the artifact itself, the artifact designer, and the artifact user, together form a moral entity that can be morally evaluated. A related argument is made by Hew (2014), who argues that 'For an artificial agent to be morally praiseworthy, its rules for behavior and the mechanisms for supplying those rules must not be supplied entirely by external humans' (ibid p. 197). He claims that this is not technologically feasible for foreseeable artificial agents, and that systems based on technologies such as machine learning, evolutionary computing and

self- organisation are all dependent on rules supplied by humans.

Johnson and Miller's argument rings true to the present author, and provides some explanation of the reasons why some writers and researchers are keen to believe that robots are, or could become, full moral agents, and why some are against the idea. In a similar way, there are those who insist that there is no in principle reason why it should not be possible one day to build robots that can feel pain and have emotions. And there are those who do not believe in this possibility. It does seem to come down to a belief, since there is certainly little in the way of tangible evidence that this will ever happen.

One reason for being skeptical about the likelihood that non-living, non-biological machines could develop a sense of morality at some point in the future is their lack of a biological substrate. A case can be made for the grounding of morality in biology. For instance, Churchland (2011) argues that morality in humans and mammals depends on their biology and is grounded in their ability to care for kith and kin; their recognition of other's psychological states; problem solving in a social context; and learning social practices. The basis for caring, she writes, lies in the neurochemistry of attachment and bonding in mammals. Humans and other mammals extend their self-maintenance and avoidance of pain in mammals to their immediate kin; feeling 'anxious and awful' when either their own well-being is threatened, or the well-being of their loved ones. They also feel pleasure when their infants are safe, and when they are in the company of others. These emotions form the basis for more complex social relationships, grounded in the rewarding pleasures of approval and belonging, and the 'generalised pain of shunning and disapproval' (Churchland 2011, p. 131), and the internalization of social standards.

An argument for a biological basis for morality implies that existing robots lack the biological basis for the development of morality. Current robots, lacking living bodies, cannot feel pain, or even care about themselves, let alone extend that concern to others. How can they empathise with a human's pain or distress if they are unable to experience either emotion? Similarly, without the ability to experience guilt or regret, how could they reflect on the effects of their actions, modify their behavior, and build their own moral framework?

How crucial are emotions and empathy for the development of morality? Docherty (2016) has argued that robots should not be allowed to make an autonomous kill decision in battle because they lack both empathy and emotion. Robots, she claims 'lack real emotions, including compassion', and 'could not truly understand the value of any human life they chose to take'. By contrast, because they possess empathy, 'people can feel the emotional weight of harming another individual', and refrain from unjustified killing. She argues that humans are able to make the judgments about proportionality that are required by the laws of war. Humans can apply judgment based on their past experience and moral consideration to assess the necessity of an attack, but Docherty thinks it unlikely that robots could be preprogrammed to do so, or that they would be able to reason about unanticipated scenarios.

Docherty (2016) describes robots and robot weapons as lacking both emotions and empathy. However Prinz (2011) questions the extent to which empathy is required for morality, and claims that empathy itself is not very motivating, and that it is subject to bias. According to Prinz, empathy is not necessary for making moral judgments, or for moral development, or for motivating moral conduct. Sentiments such as disapprobation, or emotions such as anger are more likely to form the basis for moral judgments about offensive behavior. The point is sometimes made that psychopaths lack empathy, and that they are also deficient in moral reasoning. But, as Prinz points out, psychopaths are also characterized by other emotional deficiencies, such as a low level of guilt, and an indifference to punishment, and their lack of empathy does not demonstrate its necessity for morality. Prinz's arguments rest on the careful distinctions he makes between empathy and emotions. Nonetheless, Prinz (2011) is clear that moral judgment, moral development and moral motivation do require emotions. And his discussion about the role of empathy does not refute the arguments made by Docherty (2016), since her objections to robots making life or death decision are primarily based on their general lack of emotions in general, and their inability to understand the value of human life.

Responding to the current situation

In the research we have looked at here, there is general agreement that current robots are not yet full moral agents. There is some disagreement about whether they could ever become so in the future. The situation is complicated by developments in robotics that make it increasingly possible to develop robots that look and behave in ways that create and encourage the illusion that they are able to understand and relate to humans. There is also a strong tendency to use terms to describe robots and computational agents that strongly imply that they are already ethical beings and moral agents.

Van Wynsberghe (2016) expresses her concern 'that robots are being built with at least the appearance of moral agency and that they are being placed into inherently ethical contexts', (ibid p. 313), and argues that such robots (in her case, service robots) demand ethical evaluation and reflection. She is also concerned about the use of ethically charged words (e.g. trusting, emotional attachment, socialising) to describe robots and their behaviour.

Noel Sharkey expresses similar concerns (Sharkey 2012) when he writes about the application of ethical terms to robots and machines. He talks about the way in which applying terms such as 'ethical' or 'humane' to machines leads to false attributions of abilities to them: 'They act as linguistic Trojan horses that smuggle in a rich interconnected web of human concepts that are not part of a computer system or how it operates. Once the reader has accepted a seemingly innocent Trojan term …. it opens the gate to other meanings associated with the natural language use of the term' (ibid p. 793). For instance, when Arkin writes that robots 'can be more humane in the battlefield than humans' (Arkin 2009, p. 30), his language implies that robots are capable of kindness, mercy and compassion.

Miller et al. (2016) were disturbed by the use of the word 'ethical' to describe the hole-rescuing robots of Winfield et al. (2014), and by the subsequent journalistic reporting of the study in the New Scientist by Rutkin (2014). As well as pointing out that inaccurately describing robot behaviour as ethical decision making is 'more likely to confuse than educate', (Miller et al. 2016, p. 392) they also set out some requirements for ethical decision making. They propose that ethical decision-making requires an 'openness to self-doubt' that they term an 'elenchus experience' with reference to Reed, (2013). They argue that for a machine to be considered to be capable of ethical decision-making, it needs to have '(i) a capacity to sense some aspects of the outside world (ii) an implementation of a function of merit that quantifies the acceptability of the current situation and (iii) a capacity to reprogram itself in order to improve performance in future situations.' (Miller et al. 2016, p. 393).

The elenchus experience of a machine may not be the same as a human elenchus experience. But according to Miller et al. (2016), if the machine or robot cannot reconsider a decision after it has been made in order to lead to better decision making in the future, it should not be considered to have made an ethical decision. To be described as ethical, a machine should be developing, or trying to develop, ethical expertise.

Miller et al. (2016) apply this definition of ethical decision making to the hole avoiding robots described by Winfield et al. (2014), pointing out that the robots cannot reconsider their actions given their effects, and cannot reprogram or adjust them in order to achieve a better outcome in the future. Winfield's robots have no ability to maintain a record of past decisions and their outcomes, or to reflect on these. The 'decisions' made by the robots are clearly the result of their programming, as we have already discussed. Miller et al. argue that terms such as 'ethical' should not be used without better justification to describe the apparent behaviour of a robot. Instead any description should be 'as simple as possible, making as few assumptions as possible about the capabilities of the AA [artificial agent]'(Miller et al. 2016, p. 400).

Of course, if robots are to be used in situations that impinge on humans, it is important to take steps to ensure that they will not harm them. But given these points about the misuse of ethically charged words, perhaps the robot programming undertaken by Winfield would be better described as addressing safety concerns, rather than as creating 'minimally ethical robots'. Similarly, the Andersons describe their systems as involving explicit ethical agents, but their work might be better, and more prosaically, described as being about patient reminders. In the case of Arkin's ethical governor, which involves harm to humans, perhaps his system would be better described as a military situation advisor.

An important reason for being careful about the use of language to describe the operations and underlying mechanisms of robots and computers is the need to remain aware of the unavoidable human involvement and responsibility highlighted by some (Johnson 2006; Johnson and Miller 2008; Hew 2014). As Johnson (2006) convincingly argues, even if a computational artefact is placed in a situation in which it is required to make decisions with moral consequences, the responsibility for such decisions still rests with the humans and the society that developed them and decided to deploy them there. Describing such machines as being moral, ethical, or humane, risks increasing the tendency for humans to fail to acknowledge their ultimate responsibility for the actions of these artefacts. It could encourage the use of inappropriate use of machines to make morally sensitive decisions that affect humans when they lack the moral competence that such decisions require.

An important component of undertaking a responsible approach to the deployment of robots in sensitive areas then is to avoid the careless application of words and terms used to describe human behaviour and decision making to robots. If those writing about robots were to eschew, or at least limit, the use of terms such as 'moral', 'ethical', 'humane', and 'caring' in their accounts, it would be easier to clearly assess their current abilities.

It is relevant at this point to question here the difference between the idea of robots making decisions in circumstances that require moral competence, and our increasing reliance on automatic and algorithmic decision-making. There are many crucial issues about this reliance that need to be addressed but that are beyond the scope of the present article (see Carr 2015; Susskind and Susskind 2015). However there are some important differences between the uses of robots in social roles, and the use of non-embodied computational systems. The robot in the classroom, or the robot on the battlefield, may be required to make decisions that require an understanding of the surrounding human social context. The inputs to its decision-making would be

based on the information gathered from its sensors, and on its interpretation of the social meaning of that information. This is quite different from a medical decision maker or advisor that is fed information off-line, and that does not have to rely on real-time interpretation of a social situation. Circumstances that require morally competent decision makers are those for which there is some ambiguity, and a need for a contextual understanding: situations in which judgment is required and there is not a single correct answer.

The idea that the circumstances in which morally competent decision makers are required are those in which there is some ambiguity about what the right decision is raises some questions about some of the examples we have considered here that are described as requiring *ethical* decisions. Moor (2006) described an ATM that dispenses the right amount of money as an implicit ethical agent: but is there any ambiguity here about whether or not the right amount of money should be dispensed? Likewise, when a robot prevents a proxy human robot from falling into a hole, is this an ethical, or a safety decision? Is the auto-pilot of an aeroplane making ethical decisions when it oversees a smooth take-off and landing? None of these examples involve the interpretation of a human social situation: instead they involve an understanding of the physical surroundings (in the case of hole-avoidance and flying), or of accurate data checking (in the case of the ATM). The situations that require a morally competent decision maker seem different to these.

Given these deliberations, and given what seems to be a general agreement that robots are not yet full moral agents, we turn now to a consideration of what would be the responsible way to respond. There seem to be two main alternative responses. Response 1 advocates the need to work towards the development of robots that have *some* level of ethical ability. Response 2 is to make efforts to prevent or dissuade people from deploying robots in situations and roles in which moral decisions are required. We will examine and evaluate both of these responses in turn.

Response 1: building 'ethical' robots

There are some authors who consider it both important, and possible, to develop robots and machines with some degree of ethical behaviour. For instance, Wallach (2010) advocates the building of 'moral machines' as a practical goal, motivated by 'the need to ensure that increasingly autonomous machines will not cause harm to humans and other entities worthy of moral consideration' (ibid p. 243). He suggests that artificial moral agents (AMAs) will continue to be developed for practical applications over a long period of time, and that testing these systems will enable an understanding of the limits of the implemented mechanisms. For instance, he proposes that the limitations of a system that lacks specific mechanisms such as emotions, a theory of mind, or consciousness will become apparent to engineers when they are tested and found not to be 'sufficiently sensitive to moral considerations essential for making judgments in certain situations'.

The idea that the limitations of such systems should be found by testing them seems unconvincing to the present author. Apart from anything else, the developer of an AMA is likely to be more concerned with showing its strengths than in finding its limitations. The limitations of a given system or robot, or of robots in general, are also usually easily identifiable without the need for actual testing. For example, Winfield's robots are able to prevent other robots from falling into a hole. But it would make little sense to test the robots to see if they were able to prevent other robots from, for instance, running out of energy, because that is not what they were programmed to do. It also seems unnecessary to actually build a childcare robot without phenomenal consciousness or emotions in order to demonstrate that the children left in its care for long periods start to exhibit dysfunctional behaviour and attachment problems. The question of whether or not robot Nannies are a good idea is one that can be considered and answered on the basis of knowledge about the current abilities of robots; it doesn't need a practical (and potentially risky) demonstration.

Moor (2006) was also, as we have seen, keen on the idea of developing ethical agents even if they always fall short of what is needed for a full ethical agent. He argued that it is important to examine the other types of moral agent he identified (ethical impact agents, implicit ethical agents and explicit ethical agents), and especially encouraged the development of explicit ethical agents because of the need to ensure that 'machines treat us well'. He also conjectured that programming or teaching ethics to a machine would improve our understanding of ethics.

Our understanding of ethics is indeed likely to be improved as a consequence of attempts to teach or program ethics into machines. This is also the case in many other domains of computational modeling where improvements in understanding have been gained as a result of having to be more explicit about the assumed underlying mechanisms. However, if we look at the current state of AMAs and 'minimally ethical' robots, and compare it to what is known about human moral abilities there seems to be an insurmountable gap between the two. As reported earlier, Malle (2015) provided a useful outline of what is required for moral competence (a moral vocabulary; a system of norms; moral cognition and affect; moral decision-making and action; and moral communication). There is no evidence of an artificial system that has come any way close to achieving such competence. In addition, the systems

that have been developed to date are all severely limited in scope to a particular domain.

Despite the current level of progress and achievements in this area, there are still those who continue to advocate the deployment of robots in situations in which moral decision-making will be required. For example, Arkin (2009) has argued that robots will be able to make more ethical decisions in the fog of war than humans. When it is pointed out that current systems do not have sufficient understanding of the human situation and context, the argument is sometimes made that there is no reason in principle to expect that they will not be able to develop this, and that given time, they will.

Response 2: limiting robot use

While some like to believe that at some time in the future robots and machines will become sentient, conscious, and able to understand the human world, there are others (including the present author) who prefer to focus instead on their actual capabilities. Currently existing robots are neither sentient or conscious, nor capable of understanding the complexities of social situations involving humans. They are also unlikely to become so in the near future: a statement that cannot be proved, but for which there is little convincing counter evidence. Given this, it is argued here that the responsible approach should be to identify those situations in which robots should not be deployed, and the social roles that they should not be given. This again is the contention of the present author.

There are other writers who are beginning to suggest this. Most prominent are those who are writing about the use of robots in warfare, and arguing against the deployment of lethal autonomous weapons, where the robot, machine, or weapon, makes 'decisions' about who to kill without human supervision. For example, Christof Heyns (2013), the UN Special Rapporteur on extrajudicial, summary or arbitrary executions has argued that robots should not be allowed to make lethal decisions on the battlefield, on the basis that they lack 'human judgment, common sense, appreciation of the larger picture, understanding of the intentions behind people's actions, and understanding of the values and anticipation of the direction in which events are unfolding' (2013, A/HRC/23/47).

Similar concerns have also been raised about the use of robots by the police (Sharkey 2016). Then there are authors who have looked at the use of robots for the care of older people (Sharkey and Sharkey 2012; Sparrow and Sparrow 2006; Coeckelberg 2010; Vallor 2011), and raised concerns about the extent to which robots can care for and respond to them in the way that human carers do. Likewise, concerns have been raised about the use of robots as nannies, carers, and teachers of children (Sharkey and Sharkey 2010; Sharkey 2016). In a similar vein, Sherry Turkle has written persuasively about her concerns about people developing and being encouraged to develop, relationships with computational artifacts that 'cannot love you back' (Turkle 2011).

Robots that are tasked with killing people are clearly treading on ethical territory. It is less obvious that this is the case for robots that are developed for the care and supervision of children, older people, or as companions. But how could a robot make appropriate decisions about when to praise a child, or when to restrict his or her activities, without a moral understanding? Similarly how could a robot provide good care for an older person without an understanding of their needs, and of the effects of its actions? Even a bar-tending robot might be placed in a situation in which decisions have to be made about who should or should not be served, and what is and is not acceptable behaviour. All of these seem to the present author to require both moral understanding and moral competence.

Saying that there are some situations in which robots should *not* be used is not the same as being overly negative about robot use. There are many situations in which robots can offer people something that would not otherwise be available (Sharkey 2014). The challenge is to find the right path to steer between capitalising on and benefitting from the unique opportunities that robots can offer, and avoiding a future in which robots are placed in positions and roles that require a moral understanding that they do not have.

Conclusions

In this paper, we have examined the progress made towards developing moral robots. We have seen how some have taken the route of attempting to program robots to be good. Others have proposed training, or raising, robots to develop moral understanding, or moral competence. The progress along both roads has been limited. Systems that have been either programmed or trained have so far been successfully applied only in quite narrow and specific domains.

We have considered some of the debates about the extent to which robots could ever be full moral agents. There are those who are skeptical about the possibility that robots could ever be said to be moral. Nonetheless, there are several writers (Asaro 2006; Moor 2006; and; Wallach and Allen 2009) who have looked at the possibility of developing robots, which, while not being full ethical agents, exhibit some level of ethicality. For instance, Moor distinguishes between ethical impact agents, implicit ethical agents, explicit ethical agents and full ethical agents.

The need to limit the unjustified use of terms such as moral and ethical to robots has been highlighted here. The circumstances in which morally competent decision-makers

are needed have also been discussed. In addition, in the light of the current progress, (or *lack* of progress) towards the development of robots that are full moral agents, or explicit ethical agents able to reason about, and reflect on their decisions, two responses to the current situation are identified here: (1) building 'ethical' robots and (2) limiting robot use.

Those advocating the first response who are interested in working towards the development of 'minimally ethical' robots, or explicit ethical agents, do not necessarily fall in the camp identified by Johnson and Miller (2008) as 'Computational Modellers'. Indeed, they are often motivated by the need to ensure the safety of humans as robots increasingly work near them, or with them, or are even placed in charge of them. An advantage of this response is that it is likely to advance our understanding of moral decision making in general, even if the ultimate goal of an artificial moral agent is never achieved. However it is argued here that their work could often be better described as having the goal of developing safe robots, than as developing ethical robots.

Those commending the second response of Limiting robot use, are likely to feel an affinity with others in the Computers-in-Society group identified by Johnson and Miller (ibid). Given the gap between current robot abilities, and those required for full moral agency, it is important to recognise that humans remain responsible for any deployments of robots in morally sensitive domains. Humans should not offload their responsibility for the effects of robot actions onto the robots that carry them out. It is also crucial that, recognising this responsibility, steps are taken to anticipate the potential negative effects of placing robots in situations where moral decisions are required, and that efforts are made to restrict their use. Appropriately developed and deployed robots have the potential to bring many benefits to human society, but the responsible robotics approach should have the aim of limiting their incursions into morally sensitive situations before it is too late.

References

Altmann, J., Asaro, P., Sharkey, N., & Sparrow, R. (2013). Armed Military Robots: Editorial. *Ethics and Information Technology, 15*(2), 73–76.

Anderson, M., Anderson, S., Armen, C. (2006). MedEthEx: A prototype medical ethics advisor. In *Proceedings of the eighteenth conference on innovative applications of artificial intelligence*. Menlo Park, CA: AAAI Press.

Anderson, S., & Anderson, M. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–26.

Arkin, R. C. (2007). *Governing lethal behaviour: Embedding ethics in a hybrid deliberative/reactive robot architecture*. Atlanta: Georgia Institute of Technology.

Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton, FL: CRC Press.

Asaro, P. (2012). On banning autonomous lethal systems: human rights, automation and the dehumanizing of lethal decision-making. *Special Issue on New Technologies and Warfare, International Review of the Red Cross 94*(886), 687–709.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics, 6*, 9–16.

Asimov, I., (1942, March) Runaround. *Astounding Science Fiction 29*, 94–103.

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of Biomedical Ethics*. Oxford: Oxford University Press.

Bostrom, N. (2014) *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Carr, N. (2015). *The glass cage: Where automation is taking us*. London: Bodley Head.

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Oxford: Princeton University Press.

Coeckelbergh, M. (2010). Health care, capabilities, and AI assistive technologies. *Ethical Theory and Moral Practice, 13*(2), 181–190.

Docherty, B. (2016) Losing control: The dangers of killer robots. *The Conversation*, June 16th, 2016.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines, 14*(3), 349–379.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*, 197–206.

Heyns, C. (2013). Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, A/HRC/23/47. New-York: United Nations

Johnson, D. G. (2006) Computer Systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4), 195–204

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*, 123–133.

Lin, P. (2013). The ethics of autonomous cars. *The Atlantic*, October 8th 2013. http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360.

Lin, P. (2015). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, H. Winner (Eds.), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche aspekte* (pp. 69–85). Berlin Heidelberg: Springer.

Malle, B. F. (2015). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*. DOI:10.1007/s10676-015-9367-8.

Malle, B. F., & Scheutz, M. (2014, June). Moral competence in social robots. IEEE international symposium on ethics in engineering, science, and technology. *Presented at the IEEE international symposium on ethics in engineering, science, and technology* (pp. 30–35). Chicago, IL: IEEE.

Matthias, A. (2011). Algorithmic moral control of war robots: Philosophical questions. *Law, Innovation and Technology, 3*(2), 279–301.

Miller, K. W., Wolf, M. J., & Godzinsky, F. (2016). This "ethical trap" is for roboticists, not robots: on the issue of artificial agent ethical decision-making. *Science and Engineering Ethics*. doi:10.1007/s11948-016-9785-y.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21, 18–21. doi:10.1109/MIS.2006.80.

Moor, J. H. (2007). Four kinds of ethical robot. *Philosophy Now, 72*, 12–14.

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEE Intelligent Systems, 24*, 14–20.

Prinz, J. J. (2011) Is empathy necessary for morality? In A. Coplan & P. Goldie (Eds.) *Empathy: Philosophical and psychological perspect*ives. Oxford: Oxford University Press, pp. 211–229.

Reed, R. C. (2013). Euthyphro's elenchus experience: Ethical expertise and self-knowledge. *Ethical Theory and Moral Practice, 16*, 245–259. doi:10.1007/s10677-012-9335-x.

Riedl, M.O., & Harrison, B. (2015) Using stories to teach human values to artificial agents. Paper presented at the *2nd international workshop on AI, ethics, and societ*y. http://www.aaai.org.

Ross, W. D. (1930). *The right and the good*. Oxford: Clarendon Press.

Russell, S. (2016, June) Should we fear supersmart robots? *Scientific American*, 314, 58–9.

Rutkin, A. (2014, September) Ethical trap: robot paralyzed by choice of who to save, *New Scientist*. Amsterdam: Elsevier.

Sharkey, A. (2014). Robots and human dignity: The effects of robot care on the dignity of older people. *Ethics and Information Technology, 16*(1), 53–75.

Sharkey, A. (2016). Should we welcome robot teachers? *Ethics and Information Technology, 18*(4), 283–297.

Sharkey, A. J. C., & Sharkey, N. E. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40.

Sharkey, N. (2012). The evitability of autonomous robot warfare. *International Review of the Red Cross, 94*(886), 787–799.

Sharkey, N. (2016) Policing with Robots, *Open Rights Group (Ed) 10×10: Digital rights, the next decade*.

Sharkey, N. E., & Sharkey, A. J. C. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 11*(2), 161–190.

Sorrell, T., & Draper, H. (2014). Robot carers, ethics and older people. *Ethics and Information Technology, 16*(3), 183–195.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Mind and Machine, 16*, 141–161.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*(12), 23–30.

Susskind, R., & Susskind, D. (2015) *The Future of Professions: How technology will transform the work of human experts*. Oxford: Oxford University Press.

Turkle, S. (2011). *Alone together*. New York: Basic Books.

Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. *Philosophy and Technology, 24*(3), 251–268.

Van Wynsberghe, A. (2016). Service robots, care ethics, and design. *Ethics and Information Technology, 18*(4), 311–321.

Vanderelst, D. & Winfield, A. (2016) The dark side of ethical robots, ARXiv:1606.02583v1 [cs.RO} 8 June 2016.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology, 12*, 243–250. doi:10.1007/s10676-010-9232-8.

Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski & C. Melhuish (Eds.), *Advances in autonomous robotics systems* (pp. 85–96). Berlin: Springer International Publishing.